Award Number: W81XWH-11-1-0734

TITLE: Role and Mechanism of Structural Variation in Progression of Breast Cancer

PRINCIPAL INVESTIGATOR: Ankit Malhotra

CONTRACTING ORGANIZATION: University of Virginia
Charlottesville, VA 22904

REPORT DATE: September 2012

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1 September 2012	Annual Summary	1 Sep 2011 - 31 Aug 2012

4. TITLE AND SUBTITLE

Role and Mechanism of Structural Variation in Progression of Breast Cancer

5a. CONTRACT NUMBER
W81XWH-11-1-0734

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Ankit Malhotra

E-Mail: ankit@virginia.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Virginia School of Medicine,
Charlottesville, VA 22903

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

As part of a large cohort of cancer samples we analyzed structural variants and their mechanism from a set of 12 basal type Breast invasive carcinomas (BRCA) We were able to identify a total of 1,657 high confidence breakpoints in the breast cancer datasets. Of the 154 complex mutations (CGRs), BRCA samples had 19. Although these were almost equally divided between *one-off events* and *stepwise events,* none of the BRCA samples showed evidence for chromothripsis. MMBIR and/or MMEJ are less common in tumors than in germline lineages, and CGR breakpoints are significantly depleted for microhomology relative to simple SV breakpoints.

Using ploidy-seq and next generation sequencing, on tumor sample from a 59-year-old triple negative breast cancer patient, we can study chronology of mutations in a tumor sample, and begin to get a handle on the extensive tumor heterogeneity. Preliminary data suggests that very few mutations were shared between the different sub-populations in a tumor. We also have data that suggests that in the case of this one patient sample, the genome evolved from the normal diploid state into a hypodiploid state, which then further evolved into two highly amplified states after a genome doubling event.

15. SUBJECT TERMS
Tumor evolution, complex rearrangements, mutational mechanisms, chromothripsis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	19	USAMRMC

Table of Contents

Introduction

It is widely known that cancer is a genetic disease. The DNA in tumor cells exhibits extensive structural variation in the form of insertions, deletions, inversions, translocations and focal amplifications. Some of these are causative "driver mutations" whereas others reflect genomic instability. Recent studies have examined the prevalence of such variations in breast cancer, however the role these play in early vs. late stages of breast cancer evolution, and the mechanisms by which they arise, remain relatively unknown.

With this research plan we aim to elucidate the role and mechanisms of genomic structural variation (SV) in the context of breast tumor progression. Using whole genome sequencing and computational methods we shall compare the SV profile of several tumors in different stages of cancer progression.

Body

Our hypothesis is that each tumor arises from a single somatic cell that acquires cancerous mutations, and in a population these mutations accumulate as the tumor progresses. Using the massively parallel next generation sequencing technologies (NGS), we aim to reconstruct whole genome architecture in tumor samples at a base pair resolution. Base pair resolution is important because it allows us determine the time of origin of mutations (by estimating the allele frequency of the alternate allele) and mechanism (by estimating the amount of homology at the breakpoint junction).

During the first year of the project, we based our analysis on two different yet supplementary sets of whole genome sequencing cancer datasets.

- A cohort of 64 different tumor datasets (7 distinct tumor types) from The Cancer Genome Atlas (TCGA) that included 12 basal like breast cancers tumor normal pairs. Details of the datasets, including the number of read-pairs and genomic coverage has been included in table 1.

- Three tumor subpopulations and a normal diploid from a 59-year-old triple negative (ER-, PR-, Her2) breast cancer patient who did not receive chemotherapy prior to lumpectomy. The sample was collected in collaboration with Dr. Nicholas Navin at the Texas MD Anderson cancer center, who developed an approach (Ploidy-Seq) to isolate and sequence rare tumor subpopulations from the same tumor sample. The patient's tumor was sampled from two spatially distinct regions and three subpopulations were flow-sorted by differences in DNA ploidy (in addition to the normal stromal cells). Each subpopulation was deep-sequenced at high coverage (mean 58X) on the Illumina HiSeq2000 system to identify the full spectrum of

somatic mutations, including point mutations, indels, copy number aberrations and structural variants

In the following sections we shall describe the methodologies employed in both cases and give a summary of key results achieved.

A. *Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms* (Malhotra A. et. al. - Manuscript under review at Genome Research)

Abstract: Tumor genomes are generally thought to evolve through a gradual accumulation of mutations, but the observation that extraordinarily complex rearrangements can arise through single mutational events suggests that evolution may be accelerated by punctuated changes in genome architecture. To assess the prevalence and origins of complex genomic rearrangements (CGRs), we mapped 6,179 somatic structural variation (SV) breakpoints in 64 cancer genomes from 7 tumor types (including 12 tumor/normal dataset pairs from basal-like breast tumors) and systematically screened for CGRs comprising 3 or more interconnected breakpoints. We find that CGRs are extremely common: 154 rearrangements comprise 25% of all somatic breakpoints and 75% of tumors exhibit at least one CGR. Based on copy number profiling, 63% of CGRs are consistent with originating through a single mutational event. CGRs have diverse architectures including focal breakpoint clusters, large-scale rearrangements joining breakpoint clusters from one or more chromosomes, and staggeringly complex chromothripsis events. Notably, chromothripsis is a variable phenotype amongst different tumor types, and has a significantly higher incidence in Glioblastoma samples (39%) relative to other tumor types (9%, none in breast cancers). Chromothripsis breakpoints also show significantly elevated intra-tumor allele frequencies relative to simple SVs, which indicates they arise early during tumorigenesis or confer selective advantage. Finally, assembly and analysis of 4002 somatic and 6994 germline breakpoint sequences reveals that somatic breakpoints show significantly less microhomology and this effect is stronger at CGRs than at simple SVs. These results are inconsistent with replication-based models of CGR genesis, and strongly argue that non-homologous repair of concurrently-arising DNA double-strand breaks is the predominant mechanism underlying complex rearrangements in somatic genomes.

This study includes 64 tumors from The Cancer Genome Atlas (TCGA) including 12 basal-like breast cancers (BRCA) (Details in table 1). Tumor and matched normal samples, in the form of blood or normal solid tissue (in one case both), were subjected to Illumina paired-end sequencing by TCGA.

To identify SV breakpoints we used HYDRA-MULTI, a new multi-sample version of our HYDRA paired-end mapping algorithm (Quinlan et al. 2010) that uses population-based readpair clustering (Quinlan et al. 2011). Readpairs from all 64 tumor samples and 65 normal samples were combined into a single clustering step, which enabled

simultaneous measurement of the evidence for each breakpoint in each sample. This method and several filtering steps (manuscript in review) identified 6,179 (1,657 in breast cancers) somatic rearrangement breakpoints. For simplicity, we classify breakpoints as deletions, tandem duplications, inversions, intra-chromosomal rearrangements (>1mb) and inter-chromosomal rearrangements based purely on the orientation and mapping distance of breakpoint-defining readpairs. This classification may not necessarily reflect variant type, especially at complex rearrangements. Different tumors and tumor types show different numbers and types of breakpoints (Fig. 1, with red box highlighting the breast cancer samples), as reported previously (Stratton 2011), with BRCA and LUSC samples often showing large numbers of tandem duplications, and GBM samples showing numerous large-scale rearrangements. We also identified 27,093 germline breakpoints, of which we use a high-confidence set of 9964 deletions and 1980 tandem duplications as controls in subsequent analyses.

Since DNA was not available, we used local *de novo* breakpoint assembly to assess the validation rate. We modified the SGA assembler (Simpson and Durbin 2012) to report all paths through the assembly string graph, rather than just a consensus contig. This allows for assembly of breakpoints present at relatively low (<50%) allele frequencies within tumor cell populations, as the vast majority of somatic SVs are. Contigs exhibiting split alignments consistent with the original breakpoint prediction were judged to validate the call (**Fig. 2A**). Using this method we validated 64.8% of somatic breakpoints and 58.5% of germline control breakpoints (**Fig. 2B**), with a mean contig length of 875bp (median 862bp). However, breakpoint assembly is technically difficult and may fail to produce validating contigs. For example, we were only able to assemble and validate 76.8% of the 5368 deletions that were identified by both our study and the 1000 Genomes Project (Mills et al. 2011), and validated by the latter. Assuming that 100% of the shared calls are true positives, this implies a validation rate of 84.4% for somatic breakpoints and 76.2% for germline controls, corresponding to a false discovery rate (FDR) of 15.6% and 23.8%, respectively.

There is growing evidence that a nontrivial fraction of somatic mutations are complex genomic rearrangements (CGRs) composed of multiple clustered breakpoints that cannot be explained by a single DNA end-joining or recombination event (Quinlan and Hall 2012). Cancer genome sequencing experiments have revealed highly complex genomic rearrangements involving tens to hundreds of breakpoints that appear to have arisen through a single catastrophic mutational event termed *chromothripsis* (Stephens et al. 2011). However, the true incidence of chromothripsis in cancer, and whether or not different tumor types are more or less susceptible, remain open questions. These questions have been difficult to address because studies have used different methodologies and definitions. Our work here presents us a novel opportunity to study CGR's and their prevalence and mechanisms in a large cohort of cancer samples.

To systematically identify CGRs, we developed a simple method involving two steps (Fig. 2C): 1) clustering breakpoints whose mapping positions in the reference genome are

within 100kb of one another; and 2) forming “interconnected chains” between breakpoint clusters that share calls in common, which is possible because each breakpoint represents a junction between two distinct loci in the reference genome. The end result is that all the breakpoints that comprise a complex event are interconnected and no farther than 100kb from another breakpoint in the chain.

We defined complex variants as those involving 3 or more distinct breakpoint calls. In order to minimize chain fragmentation, where distinct subsections of the same rearrangement may be reported as separate chains due to false negative breakpoint calls, we merged nearby CGRs using a distance threshold of 1mb. These methods identified 154 CGRs involving 1542 of the 6179 (25%) total breakpoints. Of these, 90 were "mild" CGRs composed of 3-4 breakpoints, 32 were "moderate" CGRs composed of 5-9 breakpoints and 32 were "extreme" CGRs composed of 10 or more breakpoints. CGRs were identified in 48 of 64 genomes (75%) representing all 7 cancer types, and are relatively evenly distributed across tumor types, with most tumors showing 1-5 CGRs (**Fig. 3D**). Thus, CGRs are detectable in most cancer genomes. To classify CGRs further we tried to determine whether the CGR was generated from a *one-off* complex mutational event, or from a series of simple mutations that occur in a *stepwise* fashion. An informative feature in this determination is the number of DNA copy number states associated with a CGR. One-off chromothripsis mutations have limited ability to generate multiple copy number states due to the limited number of chromosomes inside of a cell, and most reported chromothripsis events involve 3 or fewer states (e.g., loss, gain and unaltered). To detect CNAs, we performed circular binary segmentation (CBS) (Olshen et al. 2004) of GC-normalized read-depth measured in windows containing 5kb of uniquely mappable sequence (Quinlan et al. 2010). We refer to the junctions between adjacent genomic segments with distinct copy number as "change-points". For a CGR to be judged as potentially resulting from a one-off mutation, we required that it exhibited no more than 3 copy number states and no more than 1 amplified copy number state exceeding 4 copies, and that it was not a "focal amplification" composed of a single contiguous amplified region. These criteria are consistent with previous studies of chromothripsis and arguably more precise. Using these criteria, 97 of the 154 CGRs (63%) are consistent with being generated by a one-off mutational event. These analyses indicate that complex one-off mutations play a major role in shaping cancer genome architecture.

The 12 breast cancer samples harbored 19 CGR events (comprising of 70 somatic breakpoints), however none of the events were chromothripsis. Of the 19, 11 were generated by mild *one-off* events and 8 were generated by *stepwise* CGR events (Table 2).

There are two general repair mechanisms that are supposed to give rise to complex mutational events in genomes - a) DNA replication based template switching, or b) non-homologous or microhomology-mediated end joining following chromothripsis events. We validate each of the HYDRA-MULTI breakpoints by assembling a contig spanning the

breakpoint. This gives us the base-pair resolution at the break that enables us to calculate the amount of homology and answer the question of mechanism. Microhomology based events (MMBIR) would require homology at the breakpoint, where as end-joining would not (Fig 3A). We examined all validated high confidence somatic and germline breakpoints for homology, and as expected the somatic and germline breaks showed very different profiles of homology (Fig 3B). Exceedingly few somatic breakpoints are formed by homologous recombination: whereas 15.6% of germline breakpoints show more than 20bp of homology, this is true for only 1.1% of somatic breakpoints. This demonstrates that recombination-based mechanisms play only a very minor role in tumor genome rearrangement.

We now focus on breakpoints with little or no homology. We judge variants with 2-10bp of homology to have arisen through MMEJ or MMBIR. We judge variants with 0-1bp of homology, or a single base insertion (-1bp of alignment overlap), to result from NHEJ. Somatic breakpoints exhibit significantly less microhomology than germline breakpoints (Fig. 3B). Considering only the breakpoints with alignment overlap of -1 to 10bp, 68% of germline breakpoints show microhomology while only 56% of somatic breakpoints do, and the distributions are significantly different (MWW; p=2.06 x 10-39). Therefore, MMBIR and/or MMEJ are less common in tumors than in germline lineages. To our knowledge, this is the first demonstration of a difference in the utilization of microhomology-mediated mechanisms in germline versus somatic lineages.

To assess the role of microhomology-mediated processes in generating CGRs, we next compared the homology distribution of breakpoints from simple versus complex variants (Fig. 3B). Overall, CGR breakpoints are significantly depleted for microhomology relative to simple SV breakpoints. Whereas 57.8% of simple SV breakpoints show microhomology, only 49.2% of CGR breakpoints do, and the distribution of alignment overlap in the range of -1-10bp is significantly different between simple variants and CGRs (MWW; p=1.82 x 10-4). This demonstrates that MMBIR and/or MMEJ contribute significantly less to CGR formation than to simple somatic SVs. Taken together, our breakpoint profiling experiments reveal that the majority of CGRs detectable in tumor genomes arise through end-joining of concurrently-arising double-stranded DNA breaks, not replication-based mechanisms.

B. *Inferring Mutational Chronology in Breast Cancer by Deep-Sequencing Intratumor Subpopulations* (Wang Y.*, Malhotra A.* et. al. - Manuscript in preparation)

Abstract: Heterogeneity within a tumor sample has confounded both basic and clinical research for a long time. Here we present an approach called Ploidy-Seq that combines flow-sorting nuclei by ploidy and next-generation sequencing to identify somatic mutations in cancer patients by enriching rare tumor subpopulations. Whether it happens by a gradual accumulation of mutations throughout the life of the tumor, or by clonal expansion, each of the processes leaves a definite signature in the evolution of the mutations. By isolating the different clonal populations out of a tumor we aim to

categorize the mutational spectrum of breast cancers. The intra-tumor heterogeneity would allow us to construct evolutionary trees of tumor progression. This would also give us an insight into mutational mechanisms chronologically and allow us to answer the question whether some mechanisms occur early or late in tumor evolution.

To test out the approach we applied it on a tissue sample (T10) from a triple negative (ER-, PR-, Her2-) breast cancer patient that did not receive any chemotherapy before the lumpectomy. Based on ploidy, three sub-populations were isolated and flow-sorted using FACS. These four samples (including stromal cells) were then sequenced using the massively parallel Illumina HiSeq2000 sequencing platform. The sequence (with a mean coverage of ~58X) was then used to identify all measures of somatic variations including point mutations, indels, copy number alterations and structural variants. Although we are still working on final analysis of the sequences, we can report on some preliminary results we have obtained.

As shown in Fig 4, after micro-dissection and flow-sorting on a FACS machine, we were able to isolate four different sub-populations from the sample. Based on the differences in ploidy, we got a normal diploid population (D), two copy amplified populations (AA and AB), and one with a less than expected copy number or hypo-diploid sample (H). In order to distinguish germline from somatic mutations, we filtered all germline SNPs detected in the stroma from the tumor subpopulations. We then performed set theory operations to classify mutations as

- early (present in all subpopulations),
- intermediate (shared between two subpopulations) or
- late (exclusive to one subpopulation).

The data processing pipeline that we employed to analyze the samples has been shown in Fig 5A. We used BWA to perform the basic alignment to the hg18 reference human genome. The alignments were then sorted using samtools, and duplicates removed using Picard. The custom pipeline allows us to identify the full spectrum of somatic variation, including single nucleotide variation (SNV), Insertion/Deletions (INDEL), Copy Number Variations (CNV) and Structural Variants (SV).

Novel methodology developed specifically for this project include a split read based structural variant caller. All the sequence datasets were sequenced from an approximate ~120-150 bp insert libraries (Fig 5B) with read size of 100bp. This meant that there was significant overlap between the reads from the two ends of each segment. This limits the usefulness of traditional paired end mapping approaches, which rely on a span between the two reads to call structural variants. To overcome this limitation we came up with the split-read mapping approach (SRM). We select all reads that a) did not align back to the genome, or b) had an alignment with a soft clip greater than or equal to 25bp. Our method first attempts to merge these reads into a single contig. The expectation is that there is sufficient sequence homology amongst the two

reads to merge them into a single longer contig. We were able to merge between 45% to 65% of all the reads for the four datasets by this method. We take the merged contigs, as well as all the reads that did not merge, and align them back to the reference genome (hg18) using bwa bwasw. By using bwasw mode for alignment, we are allowing for more sensitive mappings and multiple mappings per read. The idea behind this approach is that the discordant reads / contigs have a breakpoint within the sequence, so the two parts of the same read / contig would map to two different locations i.e. the two flanking regions of the breakpoint. All the split read mappings (splitters) are then fed into a clustering algorithm, that forms clusters of overlapping splitters, and then selects the splitter with maximum support from others in the cluster as representative of the cluster. The clustering algorithm is smart enough to take in multiple datasets into a single analysis run, and calculate support for split read calls

We also ran another split read variation caller - CREST (Wang et al, 2011) on the same datasets. By using two different approaches to the same question, we get more support and cross validation for our results. We attempted assembly based validation (as described above) for all the calls made from SRM and CREST. In total we made 657 validated somatic breakpoint predictions (with no support from the diploid, T10D, population). Only 145 (22%) of these calls were shared between SRM and CREST which illustrates the power we get from using two different split read variant calling approaches

As shown in Fig. 6 (A-C), our data suggest that very few 17-49% somatic mutations are shared between all three non-diploid subpopulations. Based upon these datasets, we can hypothesize that the mutations that were shared across all the three sub-populations occurred early in the evolution of the tumor, whereas the mutations that were private to a subpopulation occurred latest. The ones shared between two sub-populations were probably steps in the evolution of the tumor from early to late. Based on this hypothesis, we can begin to create evolutionary neighbor-joining trees that illustrate our view of the progression of the disease in the patient. Fig 6 D, gives one example of such a tree composed of the different SNVs that were observed in the different cases. The tree topology suggests that the T10D evolved into T10H which then evolved into T10AA and T10AB through common ancestors n1 and n2. One of the major events in the evolution of tumor seems to be a genome-doubling event that happened just before the n2 ancestor. The genes shown were the COSMIC genes that intersected with the predicted mutations. Although this is preliminary data, we can construct evolutionary trees with the same topology using the other mutational classes as well.

We are currently in the process of finalizing the analyses from this project and writing up the manuscript for publication.

Key Research Accomplishments

- We analyzed set of 12 basal type Breast invasive carcinomas (BRCA) and found:
    - 1,657 high confidence breakpoints

  - 19 complex mutations (CGRs)
  - 11 *one-off events* and 8 *stepwise events*
  - None of the BRCA samples showed evidence for chromothripsis.
  - MMBIR and/or MMEJ are less common in tumors than in germline lineages
  - CGR breakpoints are significantly depleted for microhomology relative to simple SV breakpoints.
- Using Ploidy-Seq and next generation sequencing, we began to study the chronology of mutations in a tumor sample and the extent of tumor heterogeneity. Preliminary data suggests:
  - very few mutations were shared between the populations.
  - in this patient sample, the genome evolved from the normal diploid state into a hypodiploid state, which then further evolved into two highly amplified states after a genome doubling event.

Reportable Outcomes

- Manuscript - ***Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms****, (Malhotra A. et. al.)* is currently under a second round of review at Genome Research.
- Manuscript - ***Inferring Mutational Chronology in Breast Cancer by Deep-Sequencing Intratumor Subpopulations*** *(Wang Y.*, Malhotra A.* et. al.)* is currently in preparation.

Conclusion

We have performed a large-scale study of complex structural variation in 64 cancer genomes representing 7 tumor types. We used a new multi-sample paired-end mapping algorithm to identify 6179 somatically-acquired SV breakpoints, screened for complex genomic rearrangements, and profiled 4002 somatic and 6973 germline SV breakpoints at single base resolution. To our knowledge, we have mapped a greater number of somatic breakpoints than any study to date, and are the first to systematically map CGRs in a large set of tumor samples.

Our data indicate that complex rearrangements are an important aspect of cancer genome evolution. Three-fourths of the 64 cancer genomes showed at least one CGR, and one-quarter of all breakpoints were due to complex rearrangements. Based on copy number state profiling, 63% of CGRs are consistent with a single one-off mutational events, and these comprise 13.6% of all somatic breakpoints discovered in this study

We identified chromothripsis events in an unbiased, automated fashion and found a significantly higher incidence in GBM (38.9%) relative to the other tumor types (8.7%). This definitively shows that chromothripsis is a variable phenotype among tumor types.

Our data also provide strong evidence that complex tumor genome rearrangements are formed predominantly through end-joining, not microhomology-mediated break-induced replication (MMBIR).

References

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K. et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**(7332): 59-65.

Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., and Hall, I.M. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**(5): 623-635.

Quinlan, A.R., Boland, M.J., Leibowitz, M.L., Shumilina, S., Pehrson, S.M., Baldwin, K.K., and Hall, I.M. 2011. Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. *Cell Stem Cell* **9**(4): 366-373.

Quinlan, A.R. and Hall, I.M. 2012. Characterizing complex structural variation in germline and somatic genomes. *Trends in genetics : TIG* **28**(1): 43-53.

Simpson, J.T. and Durbin, R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**(3): 549-556.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A. et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1): 27-40.

Stratton, M.R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science* **331**(6024): 1553-1558.

Wang, J. 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nature Methods **8**, 652–654 (2011).

Figure Legends

Fig 1. HYDRA-MULTI breakpoint calls. Stacked bar graph displaying the number of SVs in each tumor, with different SV classes shown as different colors. Shown on top is the entire set of calls, below are the calls validated by assembly. In the legend, deletions, duplications and inversion calls are smaller than 1mb; "intra-chrom" refers to intra-chromosomal rearrangements larger than 1mb; "inter-chrom" refers to inter-chromosomal rearrangements.

Fig 2. *(A)* Assembly-based validation of breakpoint calls. Readpairs where one or the other read map near a breakpoint prediction were extracted and subjected to *de-novo* assembly. Contigs were then aligned to the reference genome. Split-alignments detecting breakpoints consistent with the original call were judged as validated *(B)* Table showing the validation results for different breakpoint callsets. The first row shows the somatic mutations predicted in a single tumor sample, the second shows "somatic mutations" (false positives) predicted in a single normal sample, the third shows SV calls

present in a single tumor-normal pair, the third corresponds to the germline control breakpoints, the last to germline deletions calls that were also found by 1000 Genomes. The "Deletions in 1000 Genomes" column show the percentage of deletions that were found by 1000 Genomes, defined as 50% reciprocal overlap. The last two columns show the FDR by assembly, and by assessing the number of normal specific somatic mutations. *(C)* Method for detecting CGRs. HYDRA-MULTI calls are shown as dotted lines connecting distinct loci in the reference genome (blue bar at top), with each call predicting a single novel junction in the test genome corresponding to exactly two loci in the reference. Breakpoint "clusters" are formed from breakpoints found within a specified distance (in this case 100kb) of each other in the reference genome, and complex "chains" are formed from breakpoint calls linking two or more clusters to one another. *(D)* The number of CGRs observed for each tumor. *(E)* The number of breakpoints in each tumor, broken down by complexity class as shown in the legend.

Fig 3. Breakpoint homology profiles. *(A)* Measuring homology through "alignment overlap". Upon split-alignment of breakpoint-containing contigs to the reference genome, homology is apparent by overlap between the alignments on the query contig sequence (left), while "flush" breakpoints containing no homologous DNA will have alignment overlap of approximately zero. SV breakpoints harboring small insertions or small-scale rearrangements will generally have an unaligned segment, which manifests as a negative alignment overlap value. Occasionally, negative overlap values may also be caused by misalignment due to DNA sequencing errors or reference genome assembly errors at repeats. Here, overlap values less than -1 are colored light blue, those between -1 and 1 are colored orange, and those greater than 1 are colored dark blue. *(B)* Alignment overlap at germline control breakpoints (top), simple SV breakpoints (middle) and complex SV breakpoints (bottom). To aid visualization, alignment overlap values less than -1 are shown in light blue, values between -1 and 1 are shown in orange, and values larger than 1 are shown in dark blue. Please note that the X-axis scale is irregular. Overlap is measured in 1bp increments until -30 and 30, after which it is measured by tens. All breakpoints with 100 or more bases of overlap, or -100 and fewer bases, are shown at the rightmost and leftmost bars. *(C)* An X-axis zoom of the plot shown in *(B)*.

Fig 4. Overview of the Ploidy-Seq pipeline

Fig 5. Overview of the computational pipeline.

Fig 6. *(A)* A Venn diagram showing the identified SNVs in the T10AA, *(B)* T10AB and *(C)* T10H. *(D)* Neighbor joining tree drawn using the SNVs identified from the four different subpopulations. The indicated genes are the COSMIC genes that were affected by these SNVs

Figure 1

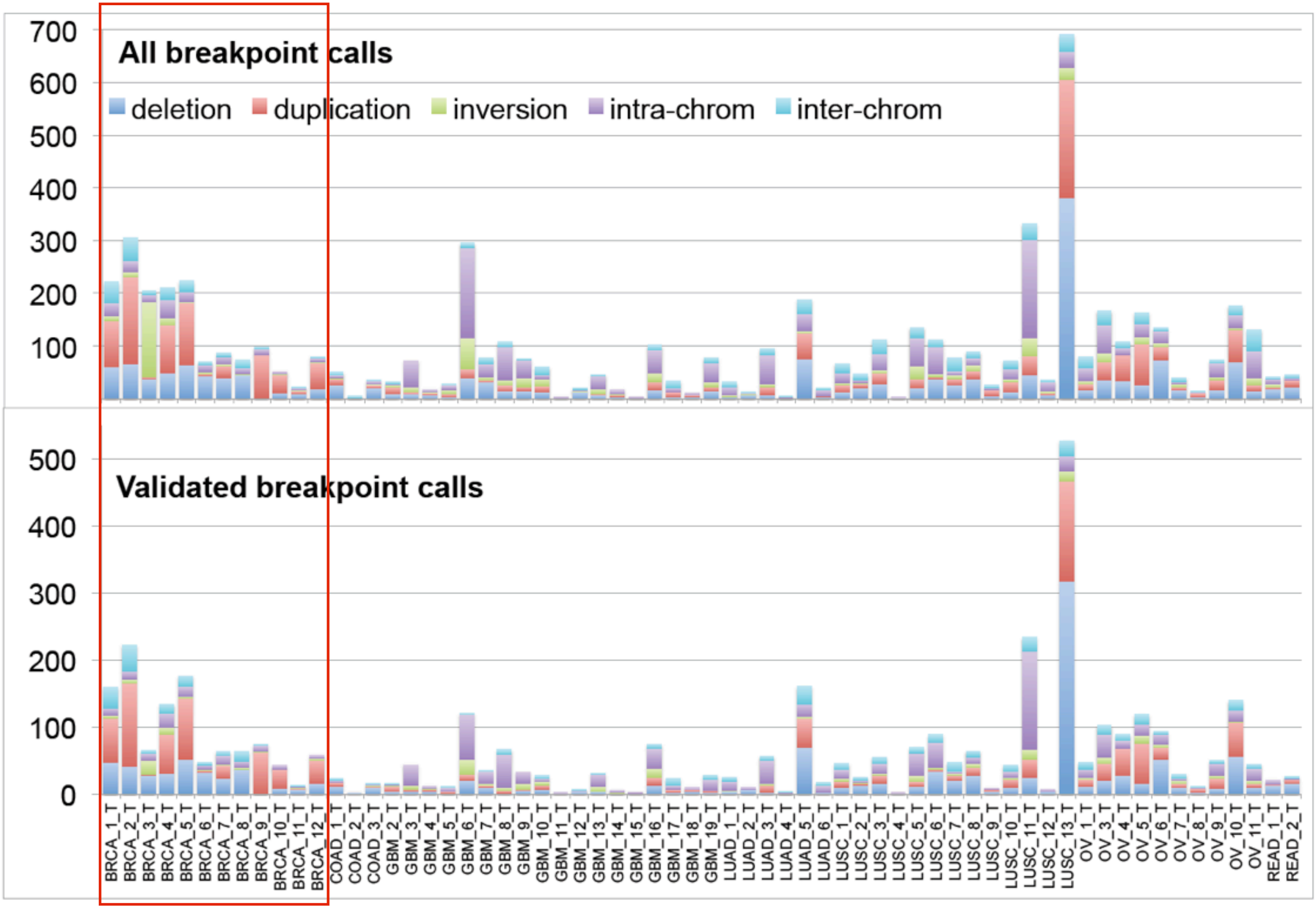
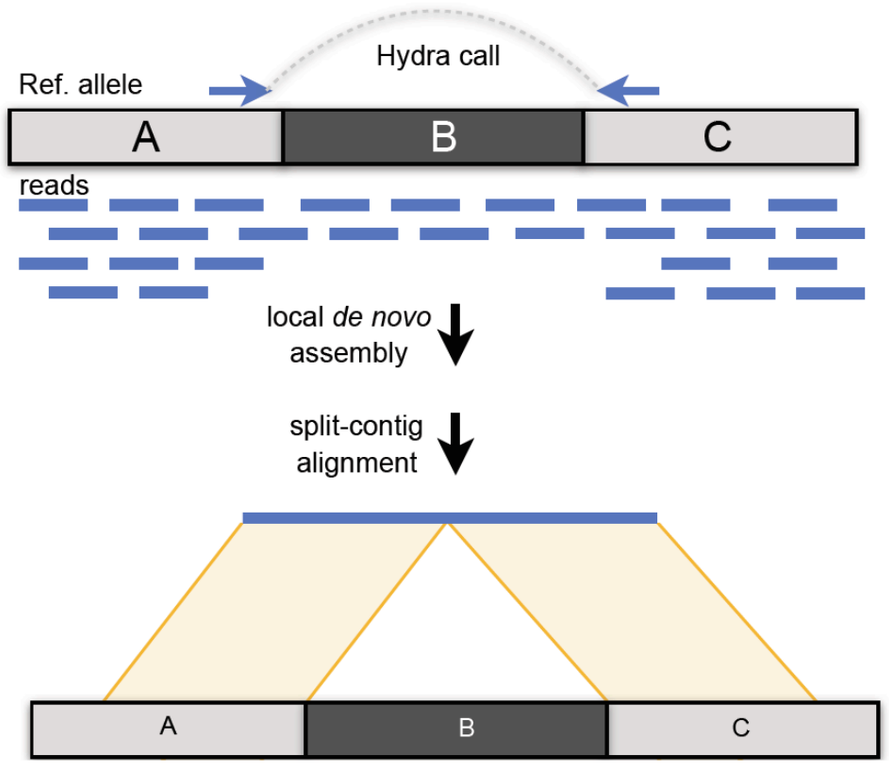

All breakpoint calls
deletion
duplication
inversion
intra-chrom
inter-chrom
700
600
500
400
300
200
100
Validated breakpoint calls
500
400
300
200
100
0
BRCA_1_T
BRCA_2_T
BRCA_3_T
BRCA_4_T
BRCA_5_T
BRCA_6_T
BRCA_7_T
BRCA_8_T
BRCA_9_T
BRCA_10_T
BRCA_11_T
BRCA_12_T
COAD_1_T
COAD_2_T
COAD_3_T
GBM_2_T
GBM_3_T
GBM_4_T
GBM_5_T
GBM_6_T
GBM_7_T
GBM_8_T
GBM_9_T
GBM_10_T
GBM_11_T
GBM_12_T
GBM_13_T
GBM_14_T
GBM_15_T
GBM_16_T
GBM_17_T
GBM_18_T
GBM_19_T
LUAD_1_T
LUAD_2_T
LUAD_3_T
LUAD_4_T
LUAD_5_T
LUAD_6_T
LUSC_1_T
LUSC_2_T
LUSC_3_T
LUSC_4_T
LUSC_5_T
LUSC_6_T
LUSC_7_T
LUSC_8_T
LUSC_9_T
LUSC_10_T
LUSC_11_T
LUSC_12_T
LUSC_13_T
OV_1_T
OV_3_T
OV_4_T
OV_5_T
OV_6_T
OV_7_T
OV_8_T
OV_9_T
OV_10_T
OV_11_T
READ_1_T
READ_2_T

Figure 2

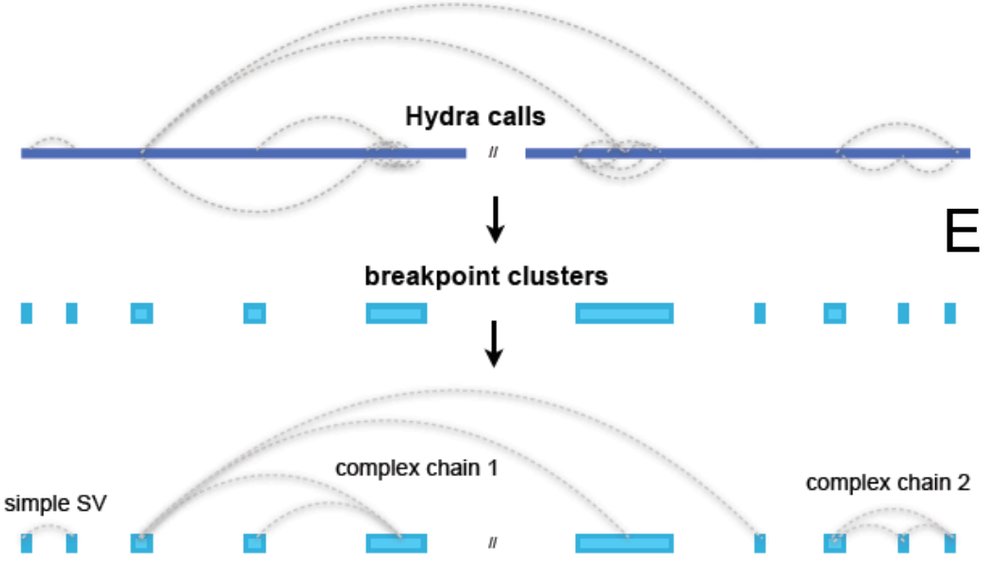
A



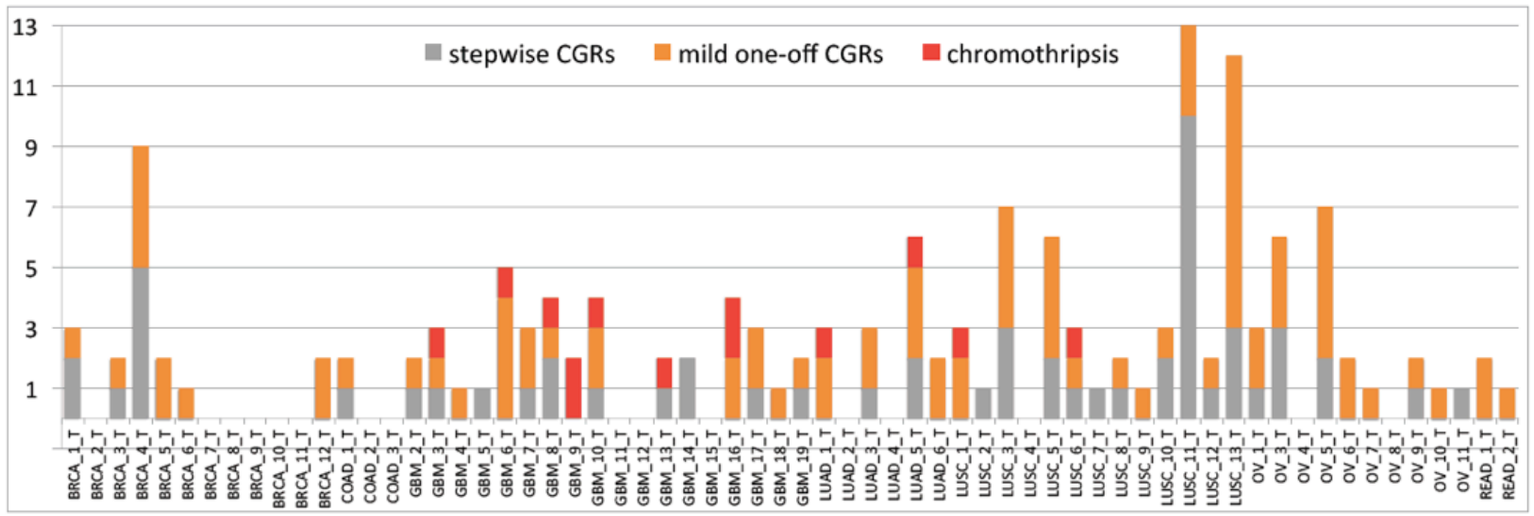
B

	num. breakpoints	validated by assembly	Deletions in 1000 Genomes	FDR by assembly (corrected)	FDR by somatic calls in normals
tumor-specific somatic mutations	6179	64.8%	2.2%	15.6%	5.2%
normal-specific somatic mutations	323	40.6%	37.0%	47.1%	-
individual-specific rare germline SVs	3297	61.9%	26.1%	19.4%	-
germline deletions & duplications	11944	58.5%	53.9%	23.8%	-
germline deletions in 1000 Genomes	5368	76.8%	100%	0	-

C



D



E

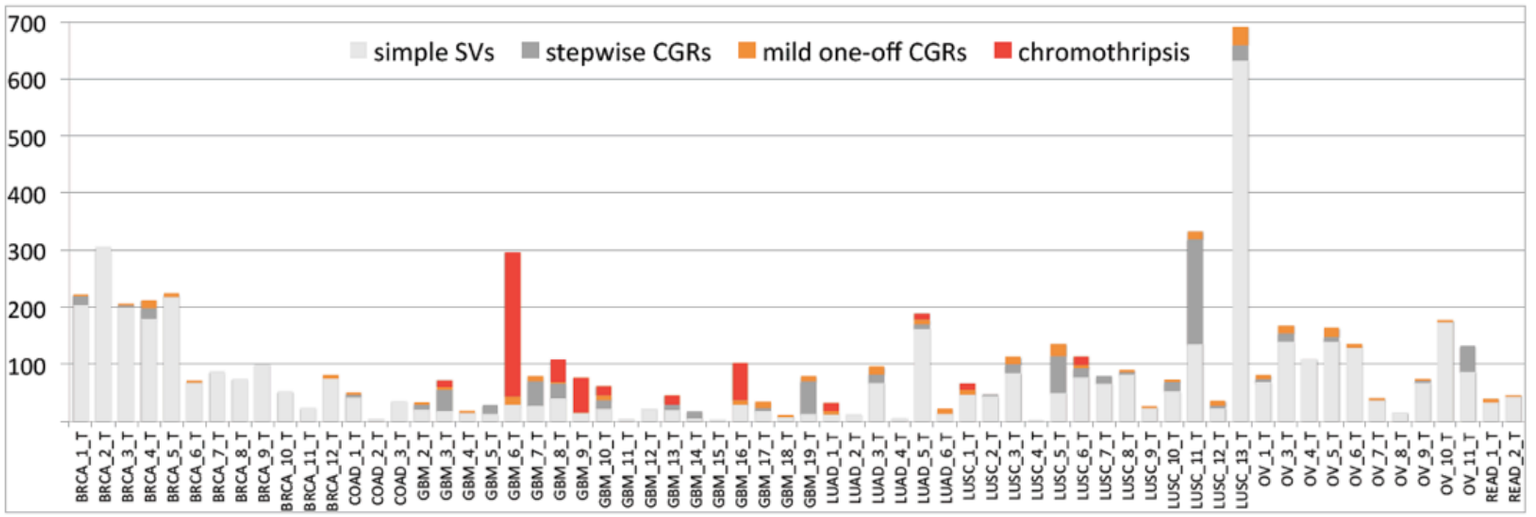
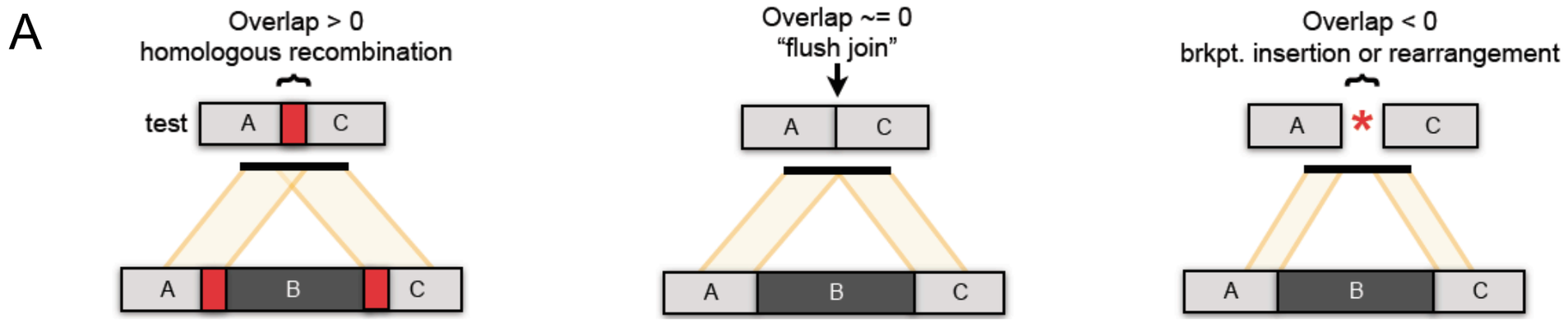
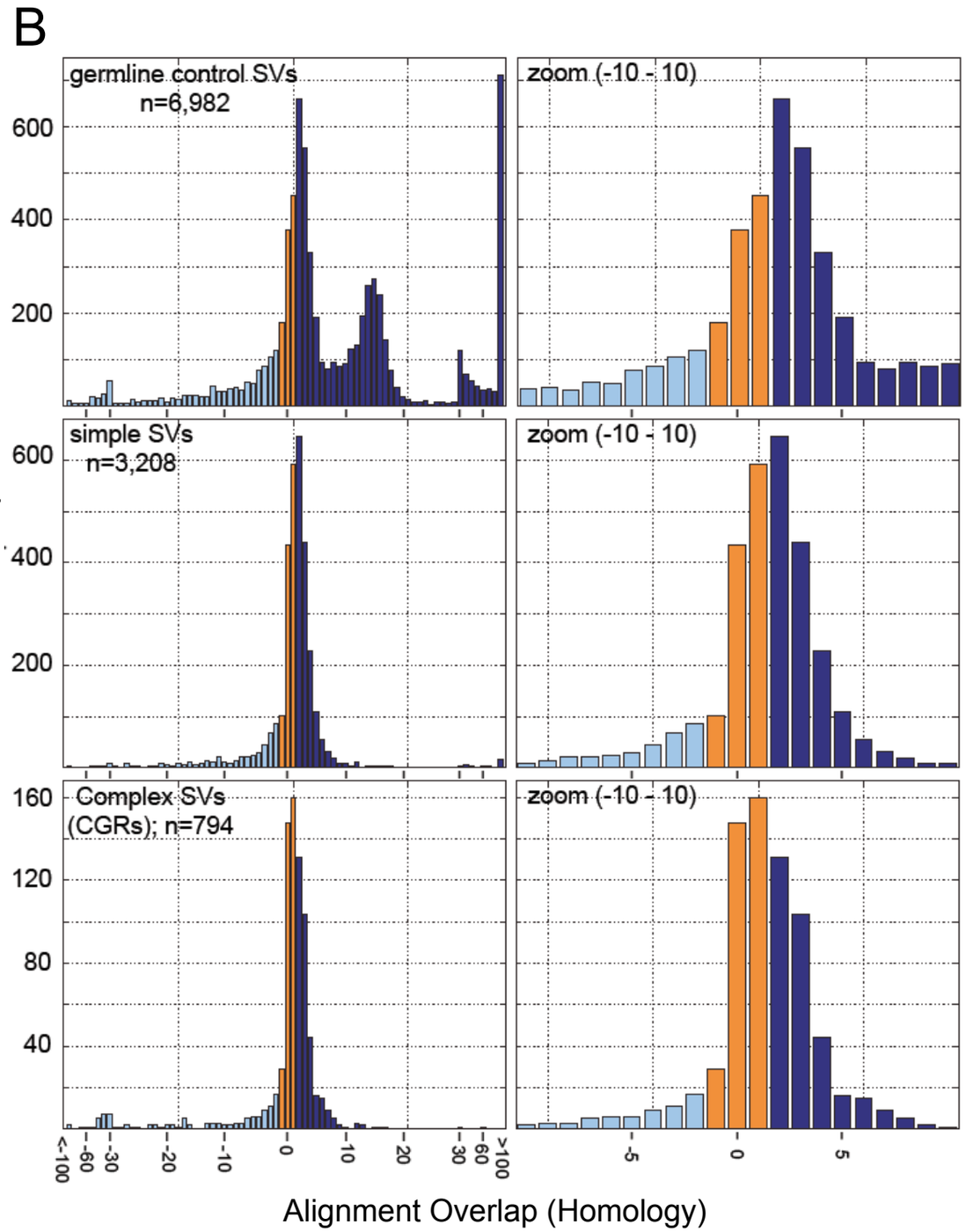
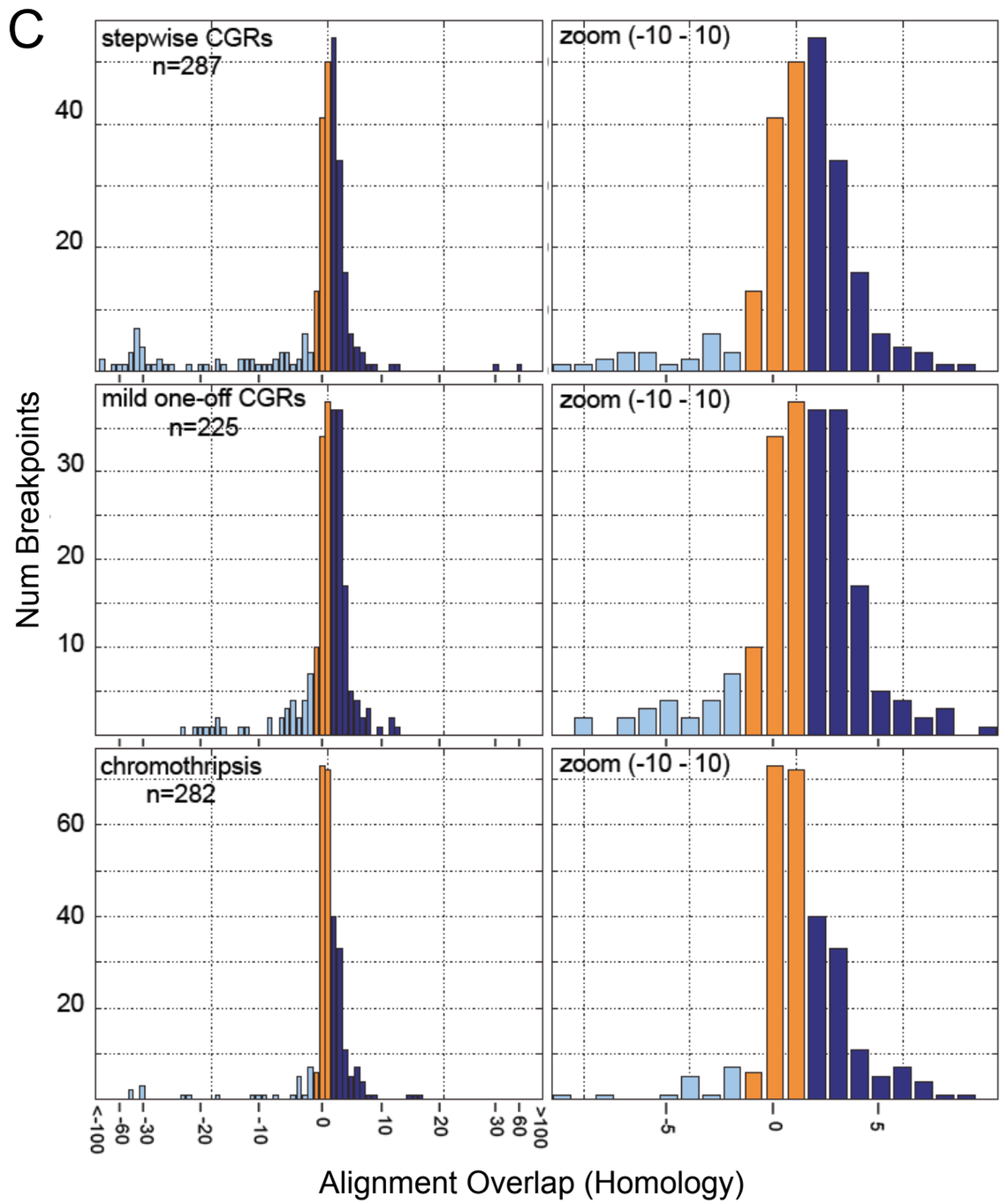


Figure 3

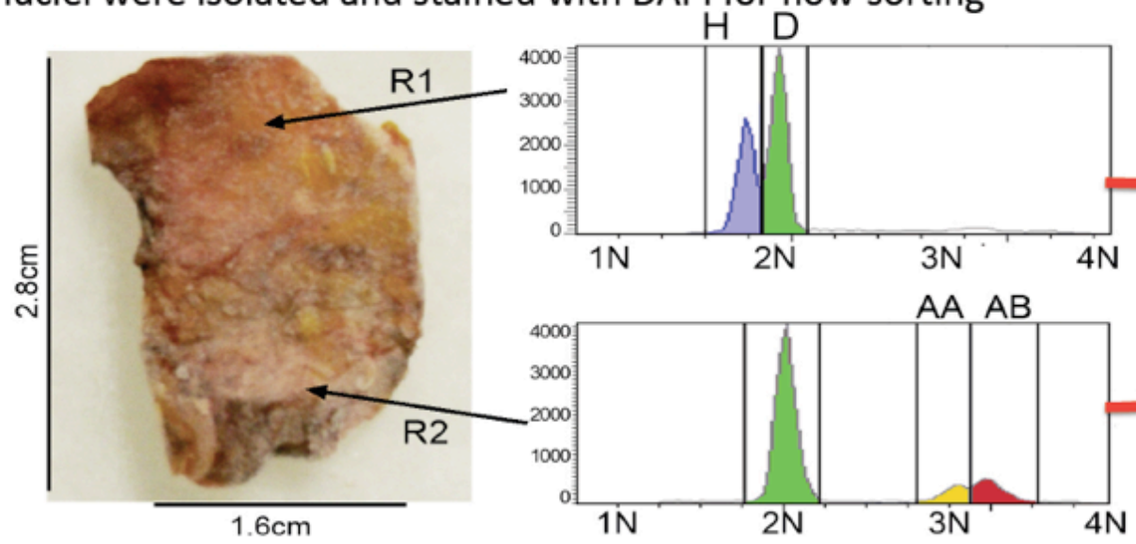

A
Overlap > 0
homologous recombination
test
A
C
A
B
C
Overlap ~= 0
"flush join"
A
C
A
B
C
Overlap < 0
brkpt. insertion or rearrangement
A
*
C
A
B
C


B
germline control SVs
n=6,982
zoom (-10 - 10)
simple SVs
n=3,208
zoom (-10 - 10)
Complex SVs
(CGRs); n=794
zoom (-10 - 10)
Num Breakpoints
Alignment Overlap (Homology)


C
stepwise CGRs
n=287
zoom (-10 - 10)
mild one-off CGRs
n=225
zoom (-10 - 10)
chromothripsis
n=282
zoom (-10 - 10)
Num Breakpoints
Alignment Overlap (Homology)

Methods : Ploidy-Seq

1. Two regions were sampled from a frozen tumor specimen. Cells were lysed and nuclei were isolated and stained with DAPI for flow-sorting



2. Four subpopulations were isolated by differences in ploidy

H: 1.75N

D – normal cells: 2N

AA: 3.1N

AB: 3.3N

3. Subpopulations were deep-sequenced (58x) on the Illumina HiSeq2000 platform

4. Somatic mutations were detected and filtered from germline SNPs

- H mutations
- D -germline mutations
- AA mutations
- AB mutations

5. Set theory operations were used to classify mutations as early, intermediate or late

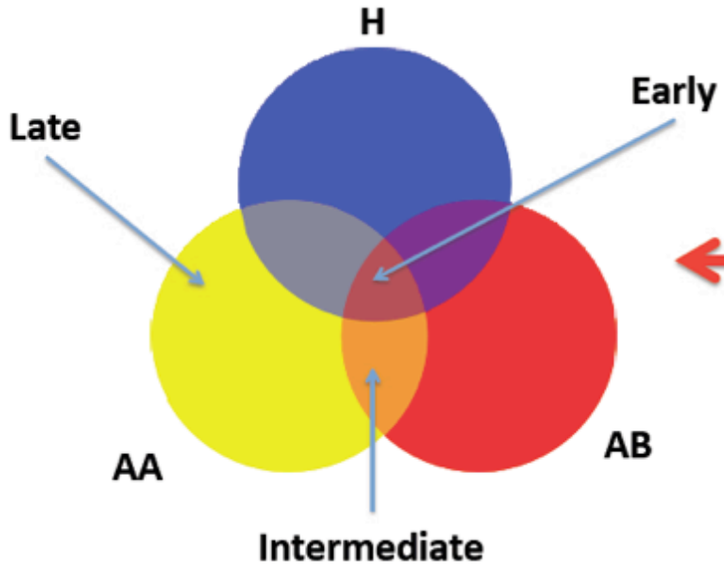
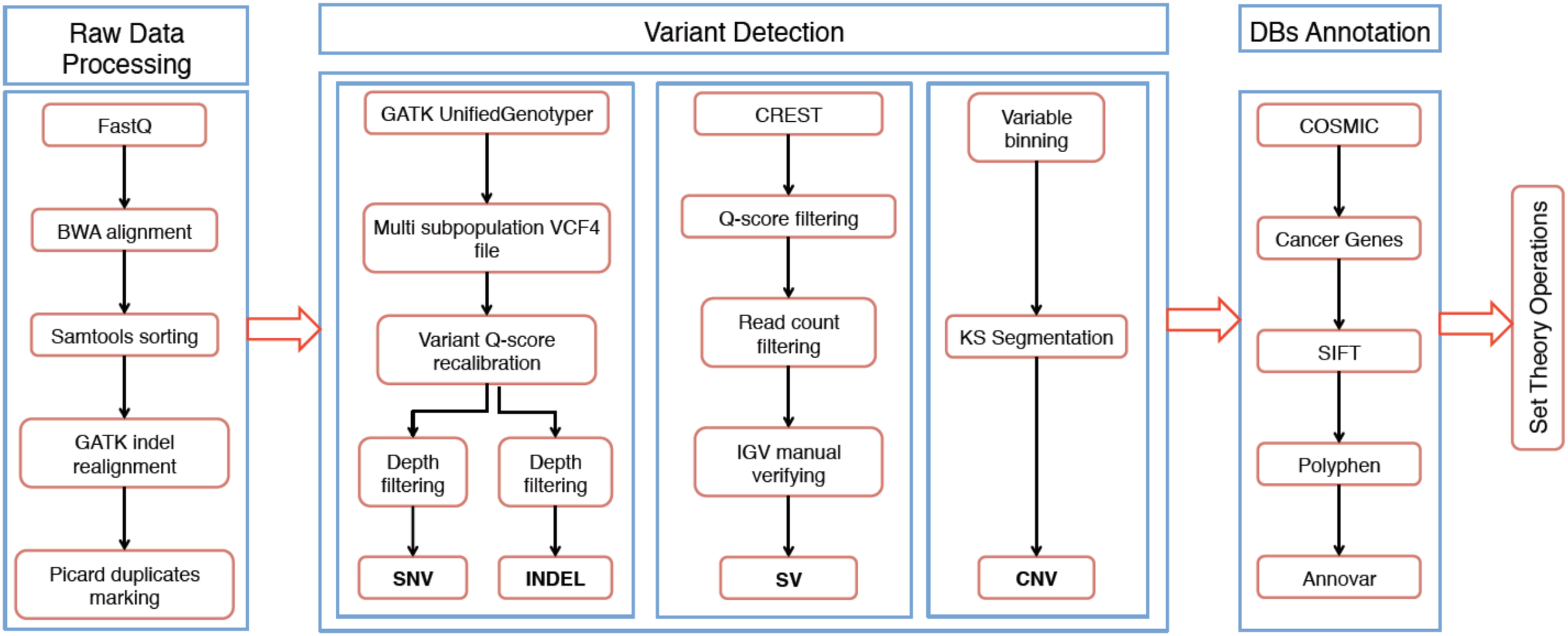


Figure 5

A


Raw Data Processing
FastQ
BWA alignment
Samtools sorting
GATK indel realignment
Picard duplicates marking
Variant Detection
GATK UnifiedGenotyper
Multi subpopulation VCF4 file
Variant Q-score recalibration
Depth filtering
Depth filtering
SNV
INDEL
CREST
Q-score filtering
Read count filtering
IGV manual verifying
SV
Variable binning
KS Segmentation
CNV
DBs Annotation
COSMIC
Cancer Genes
SIFT
Polyphen
Annovar
Set Theory Operations

B

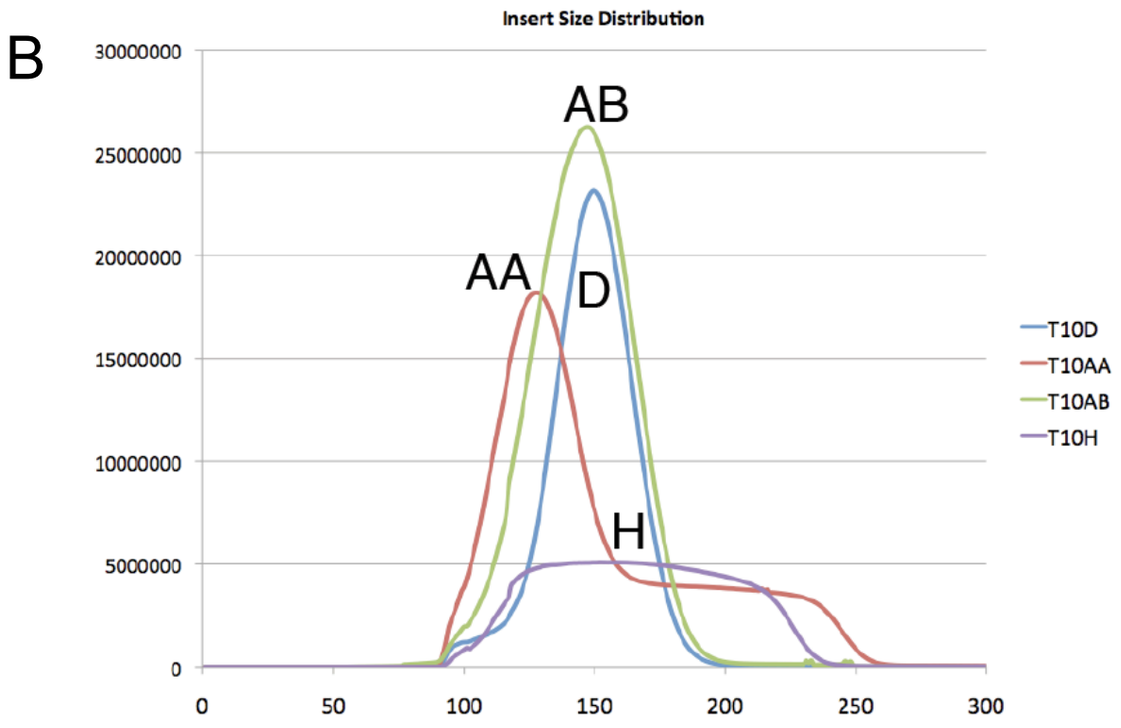
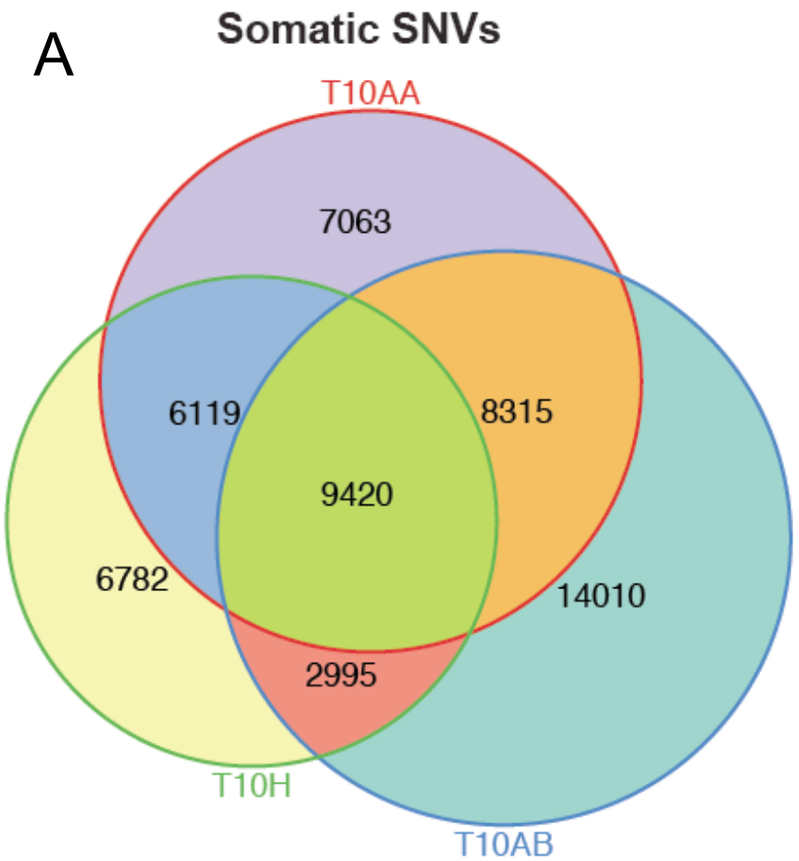
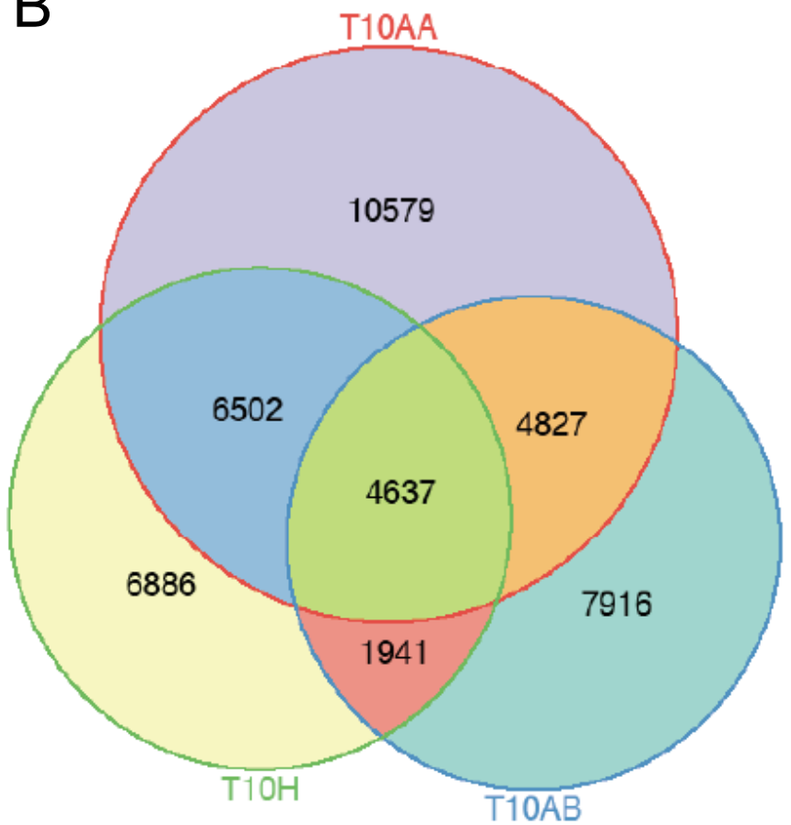

Insert Size Distribution
30000000
25000000
20000000
15000000
10000000
5000000
0
0
50
100
150
200
250
300
AB
AA
D
H
T10D
T10AA
T10AB
T10H

Figure 6

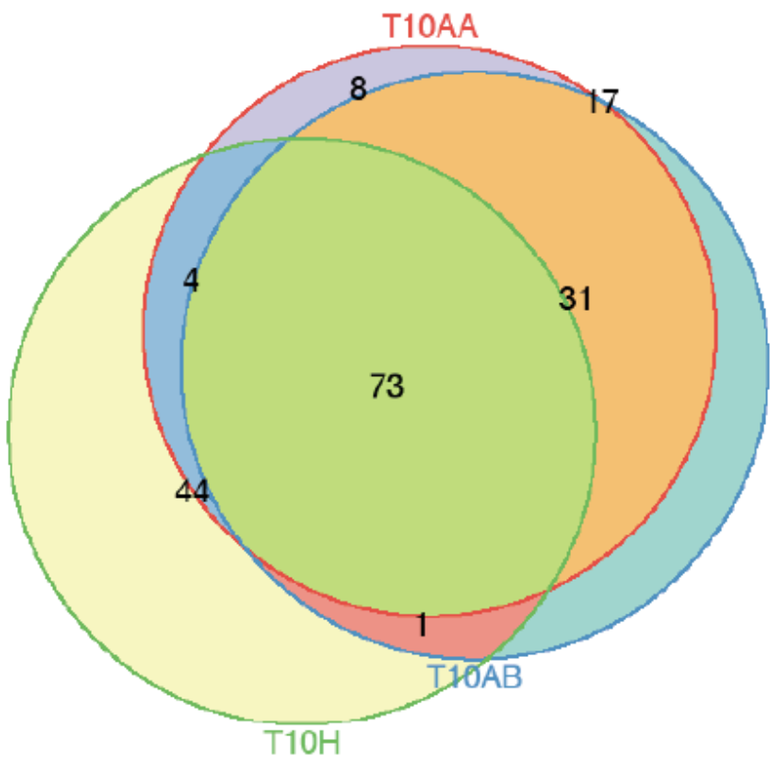
A



B



C



D